

Tipos de datos

Todos los datos existen en 2 categorias fundamentales que son:

- Categoricos: Ejemplos: Género de una persona, categoria de una película. Se clasifican así porque no son números, pueden ser cadenas de texto, pero en general son **categorías**, los datos simplemente son categorias, inclusive pueden existir categorias numericos, pero no se tratan como números.
 - Ordinales: Cuando existe una relacion de orden entre las categorias
 - Nominales: No existe una relación de orden entre las categorias.
- Numéricos: Variables que si son explicitamente numeros y debemos tratarlas como tal. La altura de una persona, temperatura.
 - Discreto: Son aquellos que tienen un valor establecido y que entre sus valores establecidos no puede existir una escala o más valores, ejemplo los números naturales que están establecidos por unidad.
 - Continuo: Son aquellos valores que pueden tener un rango amplio entre números y cuentan con valores intermedios entre ciertos valores. Un ejemplo claro es cuando usamos formato decimal, entre el 1 y el 2 hay una serie de números posibles a ocupar.

A continuación exploraremos un dataset:

<https://www.kaggle.com/lepchenkov/usedcarscatalog>

```
In [ ]: #Empecemos a trabajar con el set
#Importando pandas
import pandas as pd

In [ ]: #Definiendo un data frame para cargar el archivo
df_cars = pd.read_csv('dataset/cars.csv')
df_cars.head(3)

Out[ ]:
   manufacturer_name  model_name  transmission  color  odometer_value  year_produced  engine_fuel  engine_has_gas  engine_type  engine_ci
0         Subaru      Outback    automatic  silver      190000         2010      gasoline          False      gasoline
1         Subaru      Outback    automatic   blue      290000         2002      gasoline          False      gasoline
2         Subaru    Forester    automatic    red      402000         2001      gasoline          False      gasoline

3 rows x 30 columns
```

En este paso de cargar el Notebook, en [Deepnote](#) presenta de manera estetica la salida del Data Frame, incluso puedes escoger entre una salida con graficas para analizar ciertas variables, formato crudo (raw) o el que usa por defecto (estetico), ademas da un pequeño resumen de analisis estadistico, otra cosa interesante es que te muestra el tipo de dato que tiene cada columna del Data Frame, bastante útil es Deepnote.

¿Cómo identifico los tipos de datos en cada columna del dataset?

Eso se hace con el comando `pandas.DataFrame.dtypes` <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.dtypes.html>

```
In [ ]: df_cars.dtypes

Out[ ]:
manufacturer_name    object
model_name           object
transmission         object
color                object
odometer_value       int64
year_produced        int64
engine_fuel          object
engine_has_gas       bool
engine_type          object
engine_capacity      float64
body_type            object
has_warranty         bool
state               object
drivetrain           object
price_usd            float64
is_exchangeable      bool
location_region      object
number_of_photos     int64
up_counter           int64
feature_0            bool
feature_1            bool
feature_2            bool
feature_3            bool
feature_4            bool
feature_5            bool
feature_6            bool
feature_7            bool
feature_8            bool
feature_9            bool
duration_listed      int64
dtype: object
```

Con ese método de pandas me da inmediatamente una lista con las columnas que tiene el Data Frame y el tipo de dato.

¿Qué significa esto?

La categorización es la siguiente:

- vaiable objeto: Es una variable textual u de otro tipo, pero es variable categórica.
- bool: También se trata como una variable categórica.
- Todas las variables de tipo numérico `int64(discreto)`, `float64(contínuo)`

Con la librería de Pandas podemos generar una estadística descriptiva con varias métricas en una sola linea, usando `pandas.DataFrame.describe()` <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html>

```
In [ ]: df_cars.describe()
```

	odometer_value	year_produced	engine_capacity	price_usd	number_of_photos	up_counter	duration_listed
count	38531.000000	38531.000000	38521.000000	38531.000000	38531.000000	38531.000000	38531.000000
mean	248864.638447	2002.943734	2.055161	6639.971021	9.649062	16.306091	80.577249
std	136072.376530	8.065731	0.671178	6428.152018	6.093217	43.286933	112.826569
min	0.000000	1942.000000	0.200000	1.000000	1.000000	1.000000	0.000000
25%	158000.000000	1998.000000	1.600000	2100.000000	5.000000	2.000000	23.000000
50%	250000.000000	2003.000000	2.000000	4800.000000	8.000000	5.000000	59.000000
75%	325000.000000	2009.000000	2.300000	8990.000000	12.000000	16.000000	91.000000
max	1000000.000000	2019.000000	8.000000	50000.000000	86.000000	1861.000000	2232.000000

Me genera la tabla anterior con las columnas del data frame, pero los encabezados de fila nos están arrojando otros parametros que corresponde a una medida estadística descriptiva como:

- Count (Conteo): Me da el número de registros en el Data Frame, lo ideal es que sea el mismo en todas las columnas.
- Mean (Promedio):
- Std (Desviación Estandar):
- Min (El valor Mininmo de columna)
- 25% (Percentil 25): Cuartiles
- 50% (Percentil 50): Cuartiles
- 75% (Percentil 75): Cuartiles
- Max (El valor Máximo de columna)

Este es un resumen de las medidas que se consideran como el conjunto fundamental de estadística descriptiva, para describir un conjunto de datos.

NOTAS

Por si no sabias: JUPYTER. Son unos "cuadernos" o notebooks, en los que puedes programar por bloques. Como si escribieras un cuaderno. En una porcion, ejecutas una linea sin que tengas que correr todo el programa. Busca ANACONDA KAGGLE. Es una web que tiene concursos para analistas de datos y programadores, tiene bases de datos de uso libre. COLAB. Es el Jupyter de Google, es decir no descargas ni instalas nada, lo tienes todo en linea. DEEPNOTE. Herramienta para crear trabajos en simultaneo con otros colaboradores con los que se pueden hacer cambios en tiempo real en tus lineas de codigo [Pablo Reyes Abarca]

Extras:

<https://blog.finxter.com/pandas-cheat-sheets/> <https://cheatography.com/> <https://www.utc.fr/~jlaforet/Suppl/python-cheatsheets.pdf>